Adjusting for Survey Error in Administration Data to Create Crop Acreage Estimates From FSA Data

Michael Price (mprice@iastate.edu)
Department of Statistics, Iowa State University

Introduction

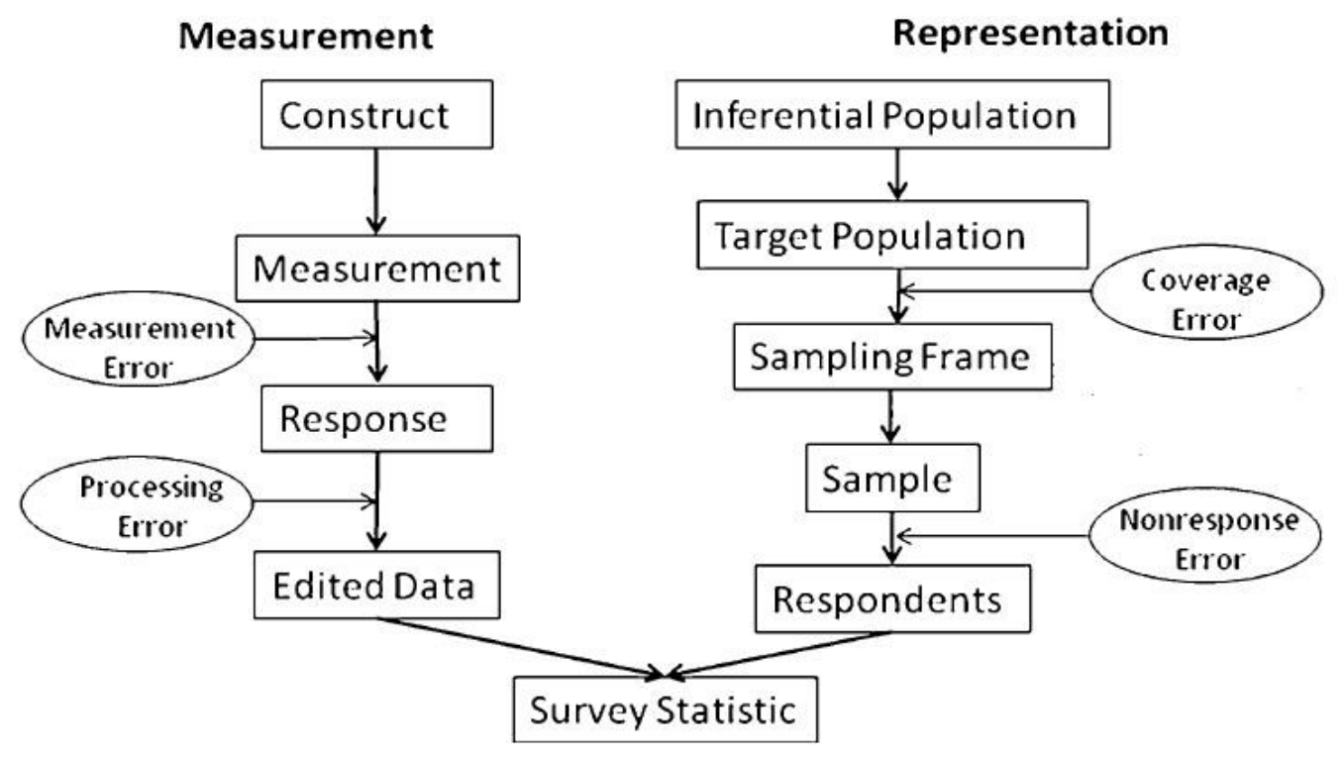
❖ Use of administrative data for statistical purposes has increased in popularity over the recent years

❖ Reasons include larger sample size and lower cost than survey data. [1]

❖ Problems and disadvantages include quality and appropriateness.

❖ Total Survey Error[2] can be used to identify issues and adjustment methods for administrative data.

The chart on the right shows Groves' chart of Total Survey Error[2] with some of the errors found in administrative data



These include:

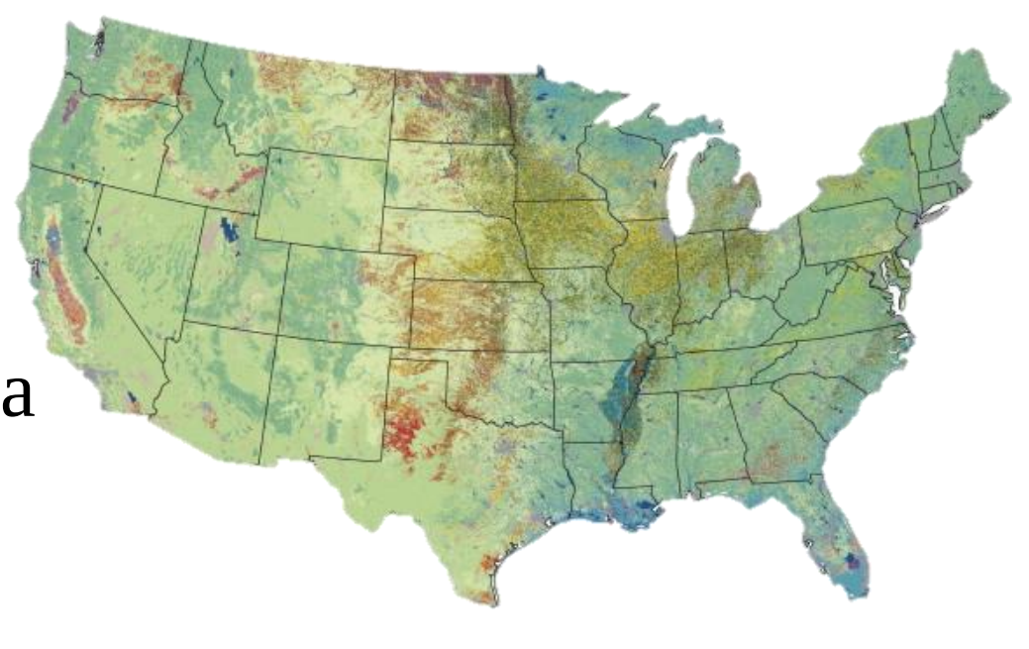
- **Coverage**-do the administrative data cover the entire population of interest?
- **Construct** -do the administrative data measure the same concept as the survey?
- **Measurement** –are there errors in the method used to collect the data?
- **Processing**-are there errors inputting or processing the data for the database?

Overall Project

Overall project goal: My research is part of a larger project to create an unbiased estimator of crop acreage through generalized method of moments using three different data sets.

The source datasets for crop acres are:

- Farm Service Agency data (FSA)-administrative data
- Cropland Data Layer (CDL)-a GIS land cover map
- June Area Survey (JAS)-sample survey data



National View of the Cropland Data Layer

Each dataset has its own pros and cons:

PROS	CONS
FSA: Contains actual planted acres Information on large number of crops for a large geographic area	**FSA:** Does not represent full universe of interest Forms of measurement and processing error
CDL: Covers entire geographic area	**CDL:** Small number of well represented crops Measurement error
JAS: Structured survey setup Information direct from the farmer	**JAS:** Small sample size High variance at county level

More specific details are discussed in Zimmer's presentation. [4]

References

1. Groen, J. (2012). Sources of Error in Survey and Administrative Data. Journal of Official Statistics, 28, 173-198.
2. Groves et al. (2004). Survey Methodology (2nd edn). Hoboken, NJ: John Wiley & Sons.
3. Parsons, J. NASS, Research Division. (1996). Estimating the Coverage of FSA Crop Acreage Totals. (SRB-96-02)
4. Zimmer, S. (2013, June). Estimates from Several Sources for Estimating Acreage of Crops. Presentation delivered at 7th international total survey error workshop , Ames, Iowa.

Advisor: Dr. Cindy Yu

Cooperative Agreement 58-3AEU-1-0012 with USDA NASS

FSA Project Goal and The Data

Goal: Provide an unbiased estimator of the amount of acres of a particular crop in a particular county using FSA data

The FSA is in charge of administering the farm programs that the United States Department of Agriculture sponsors.

They obtains acreage information on crops and fields from farmers who are part of these farm programs or in crop insurance.

Farmers who do sign up for these programs must register all of their fields, not just the fields that are part of the programs or going into insurance.

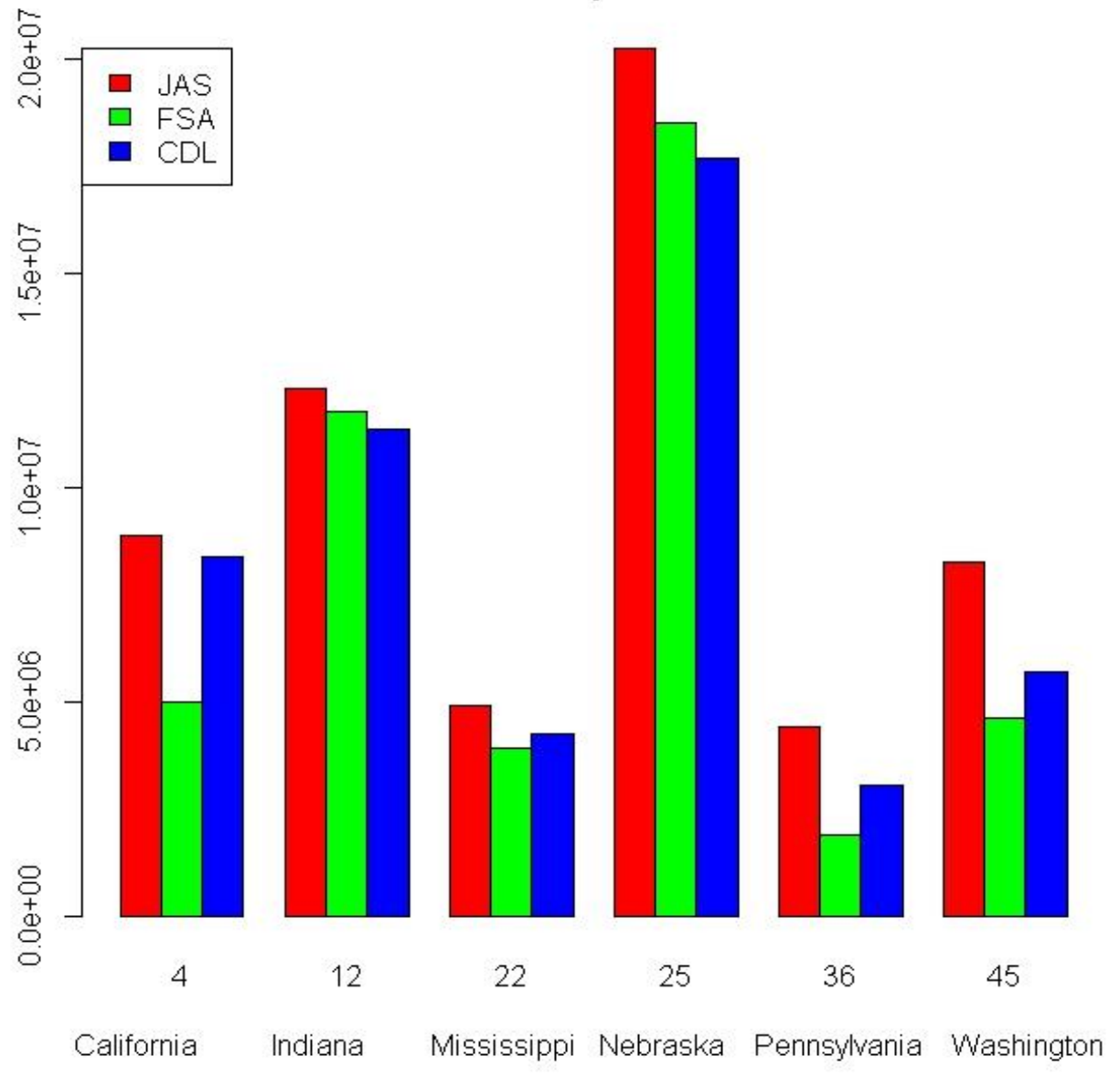
Information is stored by field in the data base. Each entry tells the field location, the amount of acres, and the crop that is insured

Survey Error In FSA Data

Biggest Problem: coverage error

❖ FSA data is not a complete set of data. Farmers may not register their fields because they:

- May not grow crops that are a part of the farm programs
- Have a small number of acres
- Do not want the government's help



Solution:

- Use ratio adjustment procedures to account for the bias introduced by the differences between the lands that are enrolled in FSA program and lands that are not
- Reference for cultivated crops acres in a county is needed
- The figure above shows the three sources of data and the amount of cultivated cropland they have measured in selected states
- Either JAS data, CDL data, or a combination of the two could be used as the reference cropland acreage
- CDL data is used because it is more accurate at the county level than JAS data

Processing Error

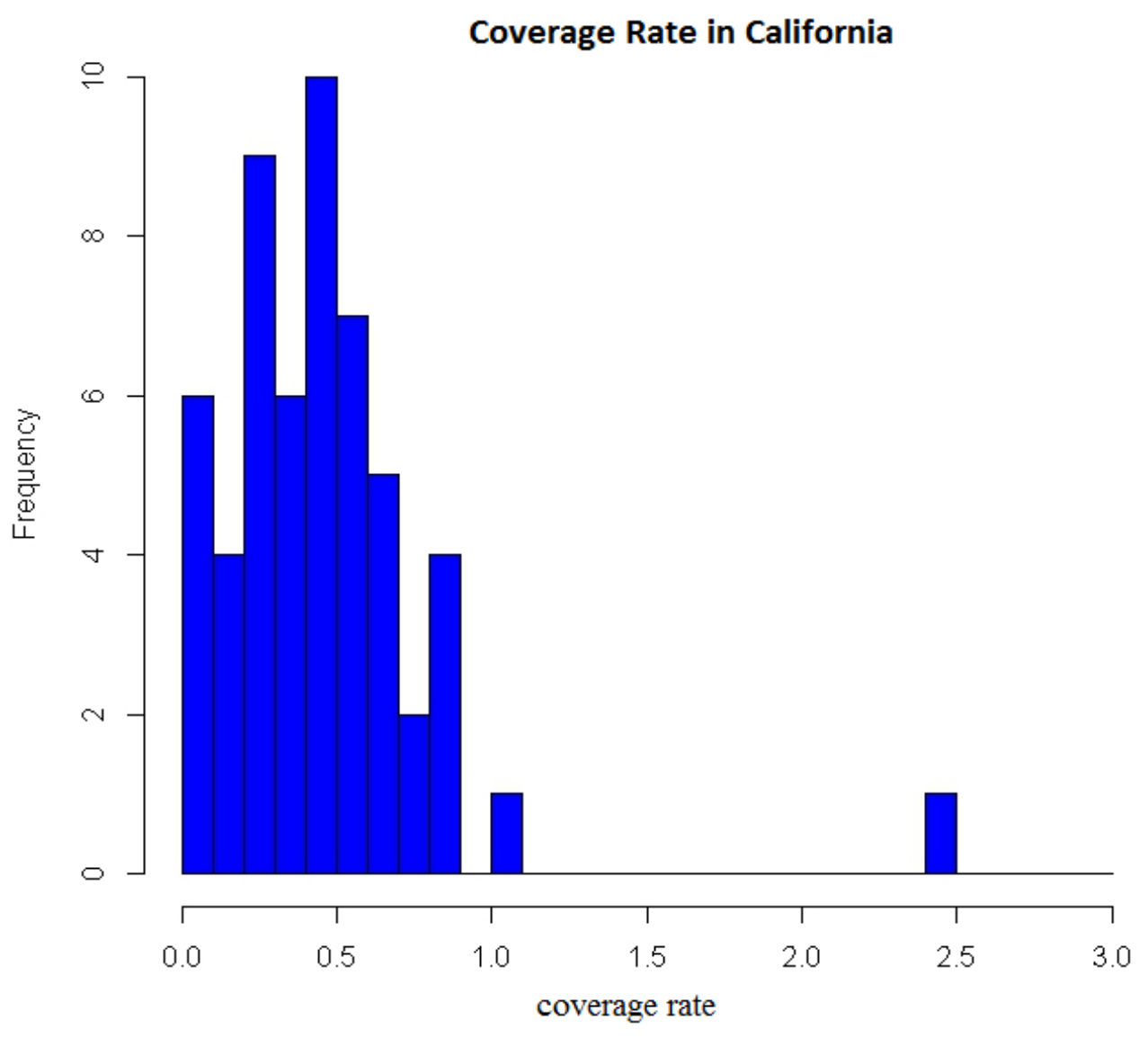
❖ The chart on the right shows county level FSA coverage rates in California

❖ Coverage rate is the proportion of cultivated land in a that is included in the FSA database

❖ This number should be less than one. Orange County is close to 2.5

WHY?

- According to the CDL, there are only 64 cultivated acres in the county
- The FSA data have one field of strawberries that is 53 acres
- It seems possible that this field is meant to be a 5.3 acre field instead
- Effect may be reduced by the proposed ratio adjustment (Estimating county acres using FSA data)



Estimating County Acres Using FSA Data

❖ To adjust for coverage error in FSA data, we estimate a propensity score for each county and then use a Horvitz-Thompson type estimator to estimate FSA acres for a state

❖ Program crops such as soybeans, wheat, corn, and cotton, are more likely to be reported than non-program crops. For example, a study showed that approximately 98% of corn and wheat fields in Nebraska were reported while only 85% of sorghum was reported[3]

❖ We assume the propensity score is homogeneous across counties in a state for program crops and homogeneous across counties for non-program crops

❖ The propensity score for county i is $\widehat{X_i}/X_i$

where $\widehat{X_i}$ is the total acres for cultivated cropland for FSA registered fields
X_i is the true value of cultivated land in the state

❖ However, the true value of cultivated land in the state is unknown so we estimate total acres for cultivated crop land from CDL data represented by $\widetilde{X_i}$

❖ To further adjust the propensity score for the difference between program and non-program crops, state level JAS crop data is used. We adjusted the propensity score so that the estimates sum to the total strata acres given by JAS

❖ This gives a final estimator of crop k as:

$$\widehat{t_{rat,ik}} = \widehat{p_h}^{-1} \times \widehat{Y_{ik}}$$

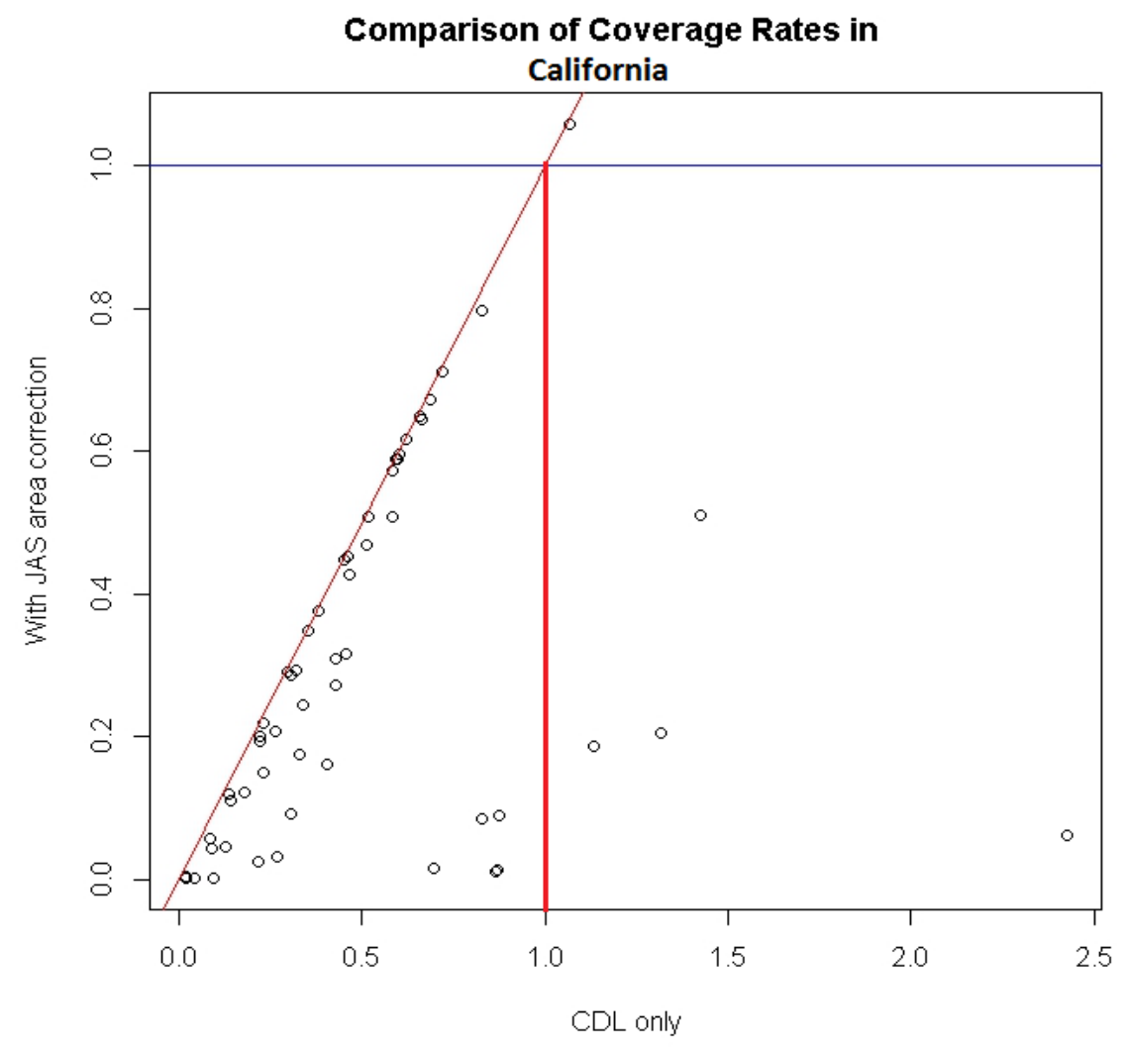
$$\widehat{p_h} = \frac{\sum_{k \in G_h} \sum_{i \in state} \widehat{\pi_i}^{-1} \widehat{t_{rat,ik}}}{\dot{Z_h}} \times \widehat{\pi}_i$$

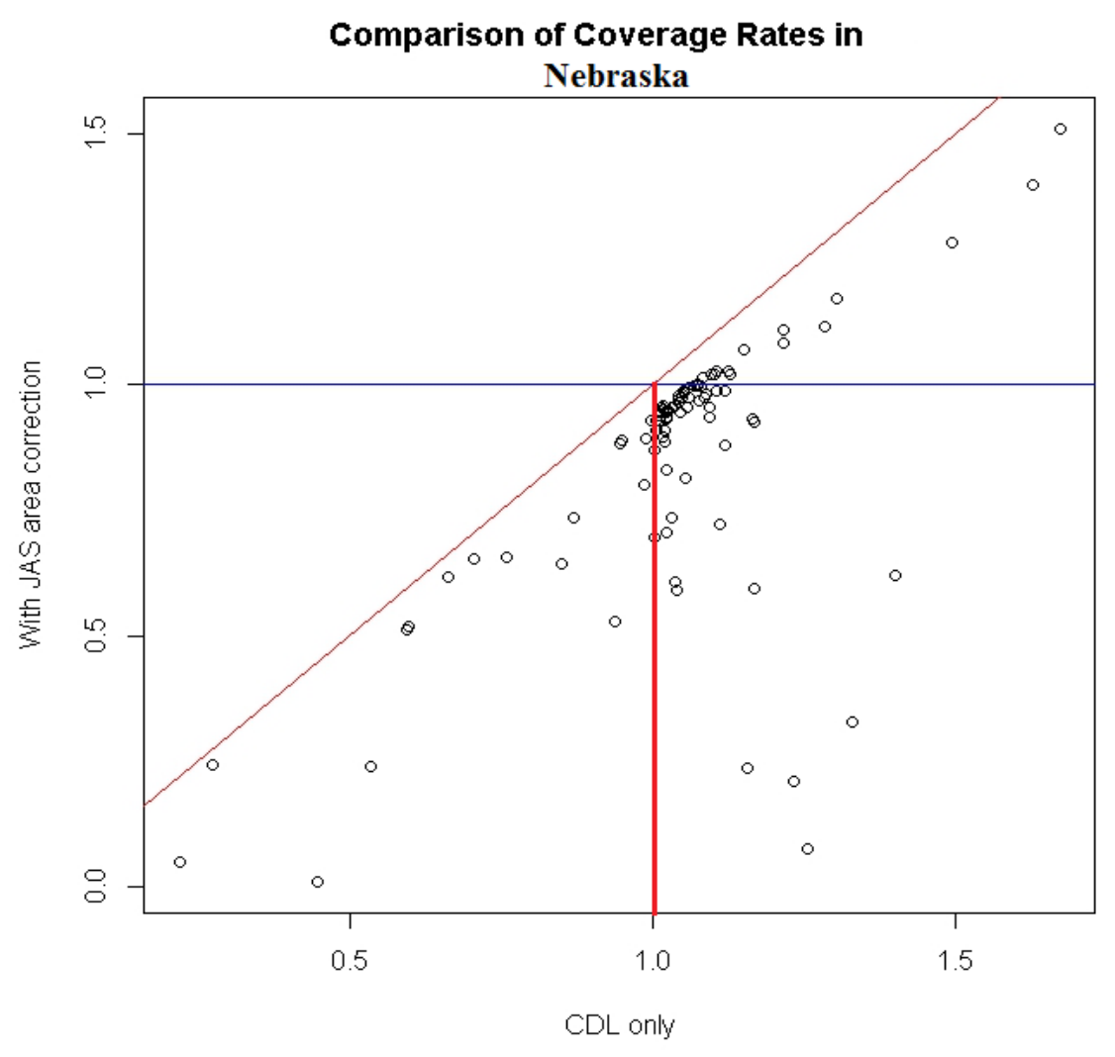
with $\dot{Z}_h$ being the state level number of acres in strata h from the JAS
and $\widehat{\pi}_i$ is $\frac{\widetilde{X_i}}{\widehat{X_i}}$

Strata are program crops and non-program crops

Adjusting for CDL Coverage Error

- We observed that many counties had propensity scores over one, represented by the dots to the right of the red line on the graphs below.





- This was believed to be due to underestimation at the county level in the CDL data source.
- To adjust for this, we calculated the difference between the unbiased JAS estimate of cultivated crop acres for a state and the estimated CDL state level cultivated acreage. This difference was then allocated to counties proportional to county land area, and then added to the CDL county acre estimate
- This drastically decreased the amount of counties with coverage rates greater than one
- All dots below the blue line have coverage rates of less than one
- It also helps adjust for the processing error that was seen in Orange County
- For the final state-level acreage estimate, all counties that still had coverage rates over one after this adjustment were truncated to one